Big data-led cancer research, applications and insights

Brown, J. A. L., Ni Chonghaile, T., Matchett, K. B., Lynam-Lennon, N., & Kiely, P. A. (2016). Big data-led cancer research, applications and insights. *Cancer Research*, *76*(21), 6167-6170. https://doi.org/10.1158/0008-5472.CAN-16-0860

Published in:
Cancer Research

Document Version:
Peer reviewed version

Title:
Big-data led cancer research, application and insights.

Authors:
James AL Brown[1 Ψ *], Triona Ni Chonghaile[2], Kyle B. Matchett[3], Niamh Lynam-Lennon[4], Patrick A Kiely[5]

[1] Discipline of Surgery, School of Medicine, The Lambe Institute, Translational Research Facility, National University of Ireland Galway, H91 YR71, Ireland.

[2] Physiology & Medical Physics Department, Royal College of Surgeons in Ireland, 123 St. Stephen's Green, Dublin 2, Ireland

[3] Centre for Cancer Research and Cell Biology, Queen's University Belfast, 97 Lisburn Road, Belfast BT9 7BL, United Kingdom

[4] Trinity Translational Medicine Institute, Department of Surgery, Trinity College Dublin, College Green, Dublin 2, Ireland

[5] Graduate Entry Medical School, University of Limerick, Limerick, Ireland

[Ψ] ORCID: 0000-0002-3155-0334

*Corresponding author: james.brown@nuigalway.ie

Abstract
Insights distilled from integrating multiple big-data or ‘omic’ datasets have revealed functional hierarchies of molecular networks driving tumorigenesis and modifiers of treatment response. Identifying these novel key regulatory and dysregulated elements is now informing personalised medicine.

Crucially, while there are many advantages to this approach, there are several key considerations to address. Here, we examine how this big-data led approach is impacting many diverse areas of cancer research, through review of the key presentations given at the Irish Association for Cancer Research meeting and importantly how the results may be applied to positively affect patient outcomes.

Introduction

Recent advances in understanding tumour development and treatment response has evolved to use a combined analysis integrating multiple parallel 'omic' datasets (genomics, proteomics, transcriptomics, microbiomics) generated from each tumour sample, producing functional hierarchies of molecular networks (Aviner *et al*, 2015; Franzosa *et al*, 2015; Kulbe *et al*, 2016; Yugi *et al*, 2016). The analysis of these functional hierarchies is enhanced by combining many of these networks into larger cohort-based (meta-) networks (targeting diseases such as ovarian, colorectal or breast cancer), creating unique insights into the mutational landscapes of cancers, the functional consequences in terms of gene expression and dysregulation of the epigenome, such as in paediatric neuro-oncology (Northcott *et al*, 2015). Generating and mapping each patients unique omic profiles onto these meta-networks will facilitate truly personalised therapeutics. This synergistic approach has also opened exciting new treatment options and facilitated the development of advanced new molecular models of disease.

Key elements of producing these molecular networks involve the integration and interpretation of multiple large-scale datasets generated from the individual omic systems of each sample (such as the genome, proteome and transcriptome). The term "big data" is used interchangeably here to cover three key aspects 1) the size and volume of information collected that must be analyzed and shared 2) the numerous types of complex information generated and processed and 3) the speed at which this data is generated (Mattmann, 2013). Specifically, the immense size of these combined data sets requires new and evolving techniques to allow their analysis, revealing highly complex links in parallel processes. Once a network for each sample is created, multiple networks are combined to reveal common key processes and molecular elements (Golden et al, 2011). Importantly, this approach requires a significant investment in computational infrastructure and bioinformatics expertise, both of which are required to effectively manage, exploit and analyse these extraordinarily large datasets. Here, we examine the application of this big data approach, through the evaluation of several key reports presented at the 2016 Irish Association for Cancer Research meeting.

Improving human tumour models through deconstruction

A key element to improving our understanding of human cancer is an accurate characterisation of the tumour microenvironment (TME). Current approaches to cancer therapeutics integrate induction of tumour cytotoxicity with modulation of the TME. A key obstacle to this approach is stratification of the TME to inform these treatment strategies and reveal potential novel treatment options (Crusz and Balkwill, 2015). To fully exploit this approach requires an in-depth understanding of the influence of genetic mutations driving the tumour and how the tumour location impacts the TME. Animal models are essential for research into the TME, but the further development of complex 3-D human cell models will complement the animal studies. Novel biophysical and biomechanical approaches are required to produce these advanced, complex human 3-D models allowing them to support the *in vitro* 3-D human tumour microenvironment in which malignant, haemopoietic and mesenchymal cells will communicate, evolve and grow.

Frances Balkwill (Barts Cancer Institute, Queen Mary University of London) leads the CANBUILD project, focusing on a group of high-grade serous ovarian cancers that metastasise to the omentum, which are frequently found at disease presentation. The ultimate aim of CANBUILD is to construct this cancerous tissue *in vitro* using autologous cells. This European Research Council and Cancer Research UK funded multi-disciplinary project is responding to an urgent need for models that facilitate examination of the interaction between human immune cells and malignant cells from the same individual, in an appropriate 3-D biomechanical microenvironment.

A key element of their approach is 'deconstruction' of the TME. This involves genomic, transcriptomic and proteomic profiling of ovarian tumours. Using big data techniques to integrate these profiles facilitates the production of a template for 'reconstruction' (defining cell types, intra- and extra- cellular signaling pathways, genetic influences). This integrated dataset will be used to produce a model that can be refined based on additional observations. Eventually it is hoped the model will provide predictions that can be tested *in vivo* and ultimately influence clinical decisions. Using the data generated from the deconstruction stage, the group is reconstructing the tumour *in silico*. This facilitates multivariate analysis of relationships between the molecular features, genes and proteins, higher order structures, tissue biomechanics, tissue architecture and cellularity

(unpublished data). The reconstruction phase is currently testing three bioengineering approaches (functionalised PEG hydrogels, peptide amphiphiles and a novel artificial omentum) to reconstruct a complex 3D TME *in vitro* (unpublished data).

Extracellular tumour signalling

It is widely appreciated that in addition to cancer cells, solid tumours contain infiltrating host cells and changes in the extracellular matrix (Egeblad et al, 2010). These infiltrating cells (including fibroblasts, endothelial cells and immune cells) have been demonstrated to fundamentally alter tumour biology, modifying key clinical parameters such as disease progression, response to therapy and metastasis (Hanahan and Weinberg, 2011). Importantly, it is currently not well understood how these underlying mechanisms facilitate the ability of tumour cells to recruit and co-opt these host cells and conversely, how these interactions modulate tumour cell function.

Multiple approaches have been used to address how tumour cells interact with local stromal cells including imaging, small molecule screening, transcriptional profiling and proteomics analysis using models ranging from complex *in vivo* to reductionist model systems (Hirata et al, 2015; Avgustinova et al, 2016; Locard-Paulet et al, 2016). A major technical challenge in dissecting cellular signalling between tumour and host cells still remains where extracting proteins from solid tumours or multicellular in vitro models typically result in loss of cell-specific information of the signalling molecules of interest.

To understand cell-specific signalling in a multicellular context Claus Jorgensen's group (Cancer Research UK Manchester Institute, The University of Manchester) have combined stable isotope labelling of individual cell populations with proteomics analysis (SILAC) (Locard-Paulet et al, 2016; Jorgensen et al, 2009; Anton et al, 2014; Tape et al, 2014). Combining the SILAC approach with a global phospho-proteomics analysis and informatics analysis of regulated phospho-motifs allows prediction of pathways that are regulated, importantly identifying the regulation in a context-dependent manner (Tape et al, 2014; Jorgensen et al., 2009).

Through these data-integrative approaches, they are combining multiple datasets in order to describe how signals are processed in a cell-specific manner. The long-term outcome will be to understand how specific signals are processed to promote tumour progression and how blocking these signals can enhance therapeutic responses.

Virally-induced epigenetic changes induce tumourigenesis

The power and insights gained by generating and combining omic data was demonstrated by John Greally (Albert Einstein School of Medicine, New York), using two examples of virally-induced neoplasia: hepatocellular carcinoma (HCC) in patients infected with hepatitis C virus, and cervical epithelial neoplasia in women infected with human papilloma virus. These tumours distinctively allow preneoplastic or early neoplastic stages of development of carcinoma to be studied, the cirrhotic and inflamed liver in HCC and the cervical intraepithelial neoplasia (CIN) stages preceding cervical carcinoma.

The hypothesis being tested was that epigenetic events create field defects predisposing to later mutational events that drive malignant transformation, using virally-induced neoplasia with distinct neoplastic stages of development. Using genome-wide DNA methylation patterns, a common theme in both tumour types was demonstrated: an early acquisition of DNA methylation that is sustained in later stages of progression, with a late event of global loss of DNA methylation coincident with carcinomatous transformation (unpublished data).

Significantly, both cancer types were characterised by a subset of loci acquiring DNA methylation at known targets of polycomb repression. An immunohistochemical study revealed that the EZH2 component of polycomb is expressed in infected cervical epithelium, with increased expression correlating with neoplasia progression (unpublished data). This work provides insights into the potential mechanisms of each of these cancer types, a foundation for the development of predictive biomarkers, and the potential targeting of polycomb for pre-neoplastic chemoprevention.

Highlighted in these studies was the need to manage and analyse the data from large numbers of individuals. Significantly, these studies also highlighted the need to collect technical metadata (data describing the experimental genomic data generation) important for correct data analysis. They found that technical influences significantly affected the DNA methylation patterns observed and these needed to be removed before high confidence findings could be identified.

Professor Greally described incorporating data from The Cancer Genome Atlas (TCGA), demonstrating the value of publically available high-quality reference datasets. Using this integrated data he described how epigenetic and transcriptional data can be combined, producing greater insights into the disease process, an integration approach that involves significant analytical and computational challenges.

Constructing the cancer specific microbiome

While there is an established symbiotic interaction between a host and their microbiota, it is only recently that clear evidence has emerged demonstrating the presence and role of microbiota in carcinogenesis (Cho and Blaser, 2012; Schwabe and Jobin, 2013; Thomas and Jobin, 2015). The use of genomic sequencing to understand and map somatic tumour mutations revealed additional non-human sequences present in many cancers, which are often filtered out prior to analysis. However, the recognition of bacteria as a key element in many human cancers highlighted the need to specifically analyse these non-human sequences. Using high throughput sequencing, followed by computational subtraction of human sequences revealed microbial in cancer samples. Susan Bullman (The Meyerson group, Department of Medical Oncology, Dana-Farber Cancer Institute) highlighted this approach. The development of computational approaches, such as PathSeq (Kostic et al, 2011), allowed the isolation of non-human DNA sequences in deep-sequenced human disease tissue samples. Subsequently, cancer-associated bacteria (such as *Helicobacter pylori*) have been described in several cancer types (de Martel *et al*, 2008; Schwabe and Jobin, 2013; Faïs *et al*, 2016). Recognizing the presence of bacteria in tumours has shed light on our understanding of cancer progression and importantly the mechanisms affecting how tumours respond to genotoxic therapeutics.

The TCGA has provided an unprecedented opportunity for sequencing-based pathogen discovery in cancer through the generation of large-scale sequencing data (up to 11,000 samples for approximately 33 human cancer types). Using an inventive computational approach, combined with the TCGA data, the Meyerson group profiled microbial signatures across more than 20 tumour types. Analyzing RNAseq and/or WGS data from more than 4,000 human tumour samples from TCGA cohorts using PathSeq allowed the identification of resident bacteria, viruses, fungi, bacteriophage and archaea within each tumour or normal tissue specimen. This led to identification of microbial species enriched in tumour tissue, compared to matched-normal tissue using LDA Effect Size analysis (Segata et al, 2011). In addition, correlations between the abundance of specific microbes and host gene expression (RNAseq data), protein profiles (RPPA data), mutation signatures (whole exome sequencing and WGS), molecular subtypes and other clinicopathological details can be used to further characterize the effects of bacteria present in tumours.

This approach identifies bacterial species that are enriched or depleted within human tumours. The application of this approach revealed an overabundance of *Fusobacterium nucleatum* in association with colorectal cancer (Kostic et al, 2012). Further demonstrating the power of this approach, DNA sequencing of cord colitis samples revealed previously unknown, non-human, sequences. The assembly of these sequences produced a draft bacterial genome with a high degree of homology to the *Bradyrhizobium* genus of bacteria, with the new strain provisionally named *Bradyrhizobium enterica*. *B. enterica* nucleotide sequences were subsequently found in biopsy specimens from cord colitis patients, but not in healthy control samples (Bhatt et al, 2013). These approaches emphasized the importance of determining differences between cancer-associated and non-cancer associated bacterial strains. Identifying different genomic features between strains and determining if these changes are shared specifically between tumour isolates could potentially reveal “high-risk” strains.

Clearly, evaluating the tumour microbiome as a key component of the tumour microenvironment will provide novel insights into how pathogens contribute to tumorigenesis, affect treatment and may ultimately lead to novel therapeutic targets.

Modelling led discovery of prognostic biomarkers

Many simpler biological systems (individual datasets) have classically been studied using statistical methods, which fail to account for time-dependent changes in functionality or network topology (arrangement of the elements). One systems biology approach to integrate changes in functionally and network topology involves the application of ordinary differential equation (ODE), used to describe how quantities change over (continuous) time, for data mining. This has allowed systems biology mathematical models to begin revealing new disease related insights.

Recent work by the Prehn Group (Centre for Systems Medicine, Royal College of Surgeons in Ireland) has implicated deregulation of the apoptosis pathway with chemotherapy resistance. Their ODE based modelling of apoptosis signalling has provided prognostic insights and predictive tools for colorectal cancer (CRC).

ODE based mechanistic mathematical models were used to predict upstream apoptosis signalling controlled by the BCL-2 family, ultimately regulating mitochondrial permeabilisation (Lindner et al, 2013). The model developed by the Prehn group proved to be a potent systems-based prognostic tool for stage III CRC and has significant potential as a predictive tool for 5-FU-based chemotherapy in stage II CRC patients. The strength of this approach was demonstrated as the model predicted patient mortality independent of pathological TNM staging and KRAS mutational status. However, predictions are strongly dependent on the previously described Consensus Molecular Subtypes (1 & 3) (Guinney et al, 2015). Current work is exploring the potential of this approach, by applying the model as a predictive prognostic tool to stratify the response of CRC patients treated with BCL-2 antagonists.

This demonstrates the power of using systems based modelling to assess complex protein-protein network interactions, producing clinically relevant prognostic tools.

Concluding remarks

The 2016 the Irish Association for Cancer Research meeting brought together some of the best international and Irish cancer researchers, highlighting the importance of big data led approaches facilitating a more complete understanding of each tumour type and helping us understand how tumours are supported within the host. A key message delivered was that comprehensively omic profiling human tumours and the surrounding tumour microenvironment can reveal mechanisms of tumorogenesis (viral and bacterial), generate prognostic biomarkers and identify potential therapeutic targets for personalised treatment.

The studies presented at this meeting highlight the need to capture datasets, generated by omic studies, describing the patient or sample. Facilitating big data management is the use of high-powered cloud computing, allowing storage, analysis through increased computing power and access (internal and external) of these large datasets (or storage in online repositories), relatively cheaply. Clearly, to exploit these combined datasets for cancer research requires accurate data management and new informative analysis processes. This will deliver on the promise of breakthroughs (fundamental and clinical) provided by *in silico* based, system wide, approaches to cancer analysis. This will allow more precise, individualised targeted therapeutic regimes to be developed and will ultimately improve patient outcomes.

Acknowledgments.

We would like to thank all speakers at the IACR meeting and apologise to those speakers whose work could not be included. We would like to acknowledge the contribution of the speakers discussed here, in writing this report.

References

Anton KA, Sinclair J, Ohoka A, Kajita M, Ishikawa S, Benz PM, Renne T, Balda M, Jorgensen C, Matter K, Fujita Y. PKA-regulated VASP phosphorylation promotes extrusion of transformed cells from the epithelium. Journal of Cell Science. 2014. 127(Pt 16):3425-33.

Avgustinova A, Iravani M, Robertson D, Fearns A, Gao Q, Klingbeil P, Hanby AM, Speirs V, Sahai E, Calvo F, Isacke CM. Tumour cell-derived Wnt7a recruits and activates fibroblasts to promote tumour aggressiveness. Nature Communications. 2016. 7:10305.

Aviner R, Shenoy A, Elroy-Stein O, Geiger T. Uncovering Hidden Layers of Cell Cycle Regulation through Integrative Multi-omic Analysis. PLoS Genetics. 2015. 11(10)

Bhatt AS1, Freeman SS, Herrera AF, Pedamallu CS, Gevers D, Duke F, Jung J, Michaud M, Walker BJ, Young S, Earl AM, Kostic AD, Ojesina AI, Hasserjian R, Ballen KK, Chen YB, Hobbs G, Antin JH, Soiffer RJ, Baden LR, Garrett WS, Hornick JL, Marty FM, Meyerson M. Sequence-based discovery of Bradyrhizobium enterica in cord colitis syndrome. New England Journal of Medicine. 2013. 369(6):517-28.

Cho I , Blaser MJ. The human microbiome: at the interface of health and disease. Nature Reviews Genetics. 2012. 13(4):260-270

Crusz SM1, Balkwill FR. Inflammation and cancer: advances and new agents. Nature Reviews Clinical Oncology. 2015. 12(10):584-96.

de Martel C, Ferlay J, Franceschi S, Vignat J, Bray F, Forman D, Plummer M. Global burden of cancers attributable to infections in 2008: A review and synthetic analysis. 2012. Lancet Oncology. 13, 607–615.

Egeblad M., Nakasone E. S., Werb Z. Tumors as Organs: Complex Tissues that Interface with the Entire Organism. Developmental Cell. 2010. 18:884–901.

Faïs T, Delmas J, Cougnoux A, Dalmasso G, Bonnet R. Targeting colorectal cancer-associated bacteria: A new area of research for personalized treatments. Gut Microbes. 2016 Mar 23:1-5.

Franzosa EA, Hsu T, Sirota-Madi A, Shafquat A, Abu-Ali G, Morgan XC, Huttenhower C. Sequencing and beyond: integrating molecular 'omics' for microbial community profiling. Nature Reviews Microbiology. 2015. 13(6):360-72. Review.

Golden A, Djorgovski S, Greally JM. Astrogenomics: big data, old problems, old solutions? Genome Biology. 2013. 14(8):129.

Guinney J, Dienstmann R, Wang X, de Reynies A, Schlicker A, Soneson C, et al. The consensus molecular subtypes of colorectal cancer. Nature medicine. 2015. 21(11):1350-6.

Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011. 144(5):646-674

Hirata E, Girotti MR, Viros A, Hooper S, Spencer-Dene B, Matsuda M, Larkin J, Marais R, Sahai E. Intravital imaging reveals how BRAF inhibition generates drug-tolerant microenvironments with high integrin β1/FAK signaling. Cancer Cell. 2015. 27(4):574-88.

Jørgensen C, Sherman A, Chen GI, Pasculescu A, Poliakov A, Hsiung M, Larsen B, Wilkinson DG, Linding R, Pawson T. Cell-specific information processing in segregating populations of Eph receptor ephrin-expressing cells. Science. 2009. 326(5959):1502-9.

Kostic AD, Ojesina AI, Pedamallu CS, Jung J, Verhaak RG, Getz G, Meyerson M. PathSeq: software to identify or discover microbes by deep sequencing of human tissue. Nature biotechnology. 2011. 29:393-396

Kostic AD1, Gevers D, Pedamallu CS, Michaud M, Duke F, Earl AM, Ojesina AI, Jung J, Bass AJ, Tabernero J, Baselga J, Liu C, Shivdasani RA, Ogino S, Birren BW, Huttenhower C, Garrett WS, Meyerson M. Genomic analysis identifies association of Fusobacterium with colorectal carcinoma. Genome research. 2012. 22:292-298.

Kulbe H, Iorio F, Chakravarty P, Milagre CS, Moore R, Thompson RG, Everitt G, Canosa M, Montoya A, Drygin D, Braicu I, Sehouli J, Saez-Rodriguez J, Cutillas PR, Balkwill FR. Integrated transcriptomic and proteomic analysis identifies protein kinase CK2 as a key signaling node in an inflammatory cytokine network in ovarian cancer cells. Oncotarget. 2016. 7(13):15648-61.

Lindner AU, Concannon CG, Boukes GJ, Cannon MD, Llambi F, Ryan D, Boland K, Kehoe J, McNamara DA, Murray F, Kay EW, Hector S, Green DR, Huber HJ, Prehn JH. Systems analysis of BCL2 protein family interactions establishes a model to predict responses to chemotherapy. Cancer Research. 2013. 73(2):519-28.

Locard-Paulet M, Lim L, Veluscek G, McMahon K, Sinclair J, van Weverwijk A, Worboys JD, Yuan Y, Isacke CM, Jørgensen C. Phosphoproteomic analysis of interacting tumor and endothelial cells identifies regulatory mechanisms of transendothelial migration. Science Signalling. 2016. 9(414):ra15

Mattmann C. Nature. 2013. 493:473–475.

Northcott PA, Pfister SM, Jones DT. Next-generation (epi)genetic drivers of childhood brain tumours and the outlook for targeted therapies. Lancet Oncology. 2015. 16(6):e293-302. Review.

Schwabe RF and Jobin C. The microbiome and cancer. Nature Reviews Cancer. 2013. 13:800–812.

Segata N1, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, Huttenhower C. Metagenomic biomarker discovery and explanation. Genome biology. 2011. 12, R60.

Tape CJ, Norrie IC, Worboys JD, Lim L, Lauffenburger DA, Jørgensen C. Cell-specific labeling enzymes for analysis of cell-cell communication in continuous co-culture. Molecular and Cellular Proteomics. 2014. 13(7):1866-76.

Tape CJ, Worboys JD, Sinclair J, Gourlay R, Vogt J, McMahon KM, Trost M, Lauffenburger DA, Lamont DJ, Jørgensen C. Reproducible automated phosphopeptide enrichment using magnetic TiO2 and Ti-IMAC. Analytical chemistry. 2014. 86(20):10296-302.

Thomas RM, Jobin C. The Microbiome and Cancer: Is the 'Oncobiome' Mirage Real? Trends in Cancer. 2015. 1(1):24-35.

Yugi K, Kubota H, Hatano A, Kuroda S. Trans-Omics: How To Reconstruct Biochemical Networks Across Multiple 'Omic' Layers. Trends in Biotechnology. 2016. 34(4):276-90. Review.

Supplemental Information.

Appendix.

Speaker	Affiliation
Prof Frances Balkwill	Barts Cancer Institute, Queen Mary University of London, UK
Dr Susan Bullman	Dana-Farber Cancer Institute and Harvard Medical School, Boston, USA
Prof Gordon J. Freeman	Dana-Farber Cancer Institute and Harvard Medical School, Boston, USA
Dr Susana Godinho	Barts Cancer Institute, Queen Mary University of London, UK
Prof John Greally	Albert Einstein College of Medicine, New York, USA
Dr Claus Jorgensen	Cancer Research UK Manchester Institute, The University of Manchester, UK
Prof Diether Lambrechts	VIB Vesalius Research Center, Leuven, Belgium
Prof Kingston Mills	Trinity College Dublin, Ireland
Prof Ciaran Morrison	National University of Ireland Galway, Ireland
Prof Jochen Prehn	Royal College of Surgeons, Dublin, Ireland
Prof Leonard Seymour	Oxford University, UK
Dr Mark Tangney	University College Cork, Ireland